Natural Language Processing - CS325
ASSIGNMENT -I
Indian Language Stemming Innovation Challenge: Beyond Porter

B.Tech. CS-AI
Semester: VIth

Submitted by:
Ridhima Tamang: 2316078
Shreya Sharma: 2316096
Shruti Nag: 2316098

INTRODUCTION

1. **Context and Importance:**
    - Stemming reduces words to their root forms, which is crucial for NLP tasks. It improves information retrieval, search engines, and text analytics.
    - It reduces vocabulary size and increases recall in document matching.
    - It serves as the backbone on which most NLP pipelines and structures including classification, clustering, and sentiment analysis are built.
2. **Challenge - Morphological Richness:NLP**
    - Indian languages are morphologically complex compared to English.
    - Nepali exhibits extensive agglutinative properties.
    - Multiple suffixes can attach to a single root word.
    - It has rich inflectional and derivational morphology.
3. **Language-Specific Focus:**
    - Chosen language: **Nepali**
    - Uses Devanagari script
    - Employs suffix-heavy morphology (unlike English prefixes)
    - Suffix patterns differ from Hindi and other Indo-Aryan languages
    - Gender, number, case, tense markers create numerous word forms
4. **Problem Statement:**
    - Generic stemmers (Porter, Snowball) designed for English Language-agnostic approaches fail on Nepali text. Over-stemming or under-stemming occurs Cannot handle Nepali-specific suffix combinations.
5. **Research Goal:**
    - Develop a Nepali-specific rule-based stemmer Encode linguistically-informed suffix stripping rules Improve accuracy over generic approaches Enable better Nepali text processing applications.

Linguistic Characteristics of Nepali:

Nepali, an Indo-Aryan language written in Devanagari script, exhibits rich morphological complexity through extensive use of suffixes that encode grammatical and semantic information.

a) Inflectional Morphology

Inflectional suffixes modify words without changing their core meaning or part of speech. Key patterns include:

- **Plural markers:** The suffix -हरू (-harū) marks plurality
  - किताब (kitāb, book) → किताबहरू (kitābharū, books)
  - बच्चा (bacchā, child) → बच्चाहरू (bacchāharū, children)
- **Case markers:** Postpositions attach to nouns to indicate grammatical relationships
  - घर (ghar, house) → घरमा (gharmā, in the house) [locative: -मा]
  - राम (Rām) → रामलाई (Rāmlāī, to Ram) [dative: -लाई]
  - पुस्तक (pustak, book) → पुस्तकको (pustakko, of the book) [genitive: -को]
  - काठमाडौं (Kāṭhmāḍauṁ) → काठमाडौंबाट (Kāṭhmāḍauṁbāṭ, from Kathmandu) [ablative: -बाट]
- **Verb tense/aspect markers:** Suffixes indicate temporal and aspectual distinctions
  - खानु (khānu, to eat) → खानेछ (khānecha, will eat) [future: -नेछ]
  - गर्नु (garnu, to do) → गर्यो (garyo, did) [past: -यो]
  - पढ्नु (paḍhnu, to read) → पढ्दै (paḍhdai, reading) [progressive: -दै]

b) Derivational Morphology

Derivational suffixes create new words, often changing part of speech:

- **Nominalization:** -ता (-tā) and -पन (-pan) form abstract nouns from adjectives
  - सुन्दर (sundar, beautiful) → सुन्दरता (sundartā, beauty)
  - मीठो (mīṭho, sweet) → मिठास (miṭhās, sweetness)
- **Agentive/occupational:** -दार (-dār), -क (-k) denote doers or professionals
  - सेवा (sevā, service) → सेवक (sevak, servant)
  - दुकान (dukān, shop) → दुकानदार (dukāndār, shopkeeper)

These morphological patterns are **language-specific** and cannot be adequately handled by generic English-oriented stemmers, necessitating tailored rule-based approaches for Nepali.

Critical Analysis of Generic Stemmer

Generic stemmers like Porter or Snowball use character-level rules designed for English. They don't understand Nepali grammar or Devanagari script patterns, so they treat Nepali suffixes as random character sequences rather than meaningful morphological markers.

Key limitations:

- No knowledge of Nepali suffix boundaries (-हरू, -लाई, -मा, -को)
- Cannot distinguish suffixes from root word components
- Apply irrelevant English-centric rules to Nepali text

Performance Analysis

Word	Correct Stem	Generic Output	Error Type	Reason
किताबहरू	किताब	किताबहरू	Under-stemming	Plural suffix not recognized
किताबलाई	किताब	किताबलाई	Under-stemming	Dative case marker ignored
किताबको	किताब	किताबको	Under-stemming	Genitive suffix not handled
घरहरू	घर	घरहरू	Under-stemming	Plural marker unrecognized
घरको	घर	घरको	Under-stemming	Genitive case marker retained
लेख्यो	लेख	लेख्यो	No stemming	Past tense marker not removed
लेख्नेछ	लेख	लेख्नेछ	No stemming	Future tense suffix ignored
बोल्दै	बोल्दै	बोल्दै	No stemming	Continuous aspect marker retained
हिँड्नेछ	हिँड	हिँड्नेछ	No stemming	Future tense suffix unrecognized
सफलता	सफल	सफलता	Under-stemming	Abstract noun suffix not stripped

गरिबी	गरिब	गरिबी	Under-stemming	Derivational suffix retained
विद्यार्थीहरू	विद्यार्थी	विद्यार्थीहरू	Under-stemming	Plural suffix not handled
विद्यार्थीलाई	विद्यार्थी	विद्यार्थीलाई	Under-stemming	Dative marker ignored
नेपाललाई	नेपाल	नेपाललाई	Under-stemming	Dative case suffix unrecognized
फूलको	फूल	फूलको	Under-stemming	Genitive marker not removed

Error types:

- **Under-stemming:** Fails to remove valid suffixes, treating variants as different words
- **No stemming:** Returns word unchanged despite removable suffixes

This analysis demonstrates that generic stemmers fail to handle Nepali morphology, justifying the need for a custom language-specific stemmer.

Design & Implementation of Custom Nepali Stemmer

The proposed Nepali stemmer employs a hybrid rule-based approach with validation mechanisms to balance suffix removal and root preservation.

Core Design Components:

- **Morphological Suffix Categories:**
  - Case markers: -लाई, -को, -की, -मा, -बाट, -सँग
  - Plural markers: -हरू, -हरु
  - Verb suffixes: -यो, -छ, -दै, -नेछ, -एको, -एकी
  - Derivational suffixes: -ता, -पन, -दार, -ई
- **Priority-Based Stripping:** Suffixes are removed using a greedy longest-match-first strategy. Longer, more specific suffixes (e.g., -हरूलाई) are checked before shorter ones (e.g., -लाई) to handle compound morphology correctly.

- **Validation Mechanisms:**
    - Minimum stem length: Prevents excessive truncation (typically 2-3 characters)
    - Edit distance constraints: Ensures stemmed form remains similar to original
    - Stopword preservation: Function words remain unchanged to maintain grammatical integrity
- **Stopword Handling:** A predefined list of Nepali stopwords (pronouns, conjunctions, particles) bypasses stemming to preserve essential grammatical elements.
- **Sandhi Rule Application:** Post-excision processing applies reverse phonological rules to address morphophonemic alternations occurring at morpheme boundaries. This includes consonant cluster simplification and vowel harmony restoration, accounting for systematic phonological changes in Nepali morphological processes.
- **Optimal Candidate Selection:** When multiple valid stems are generated, the system employs a scoring algorithm that prioritizes stems derived through removal of morphologically significant suffixes. This selection process balances suffix importance with validation metrics to identify the optimal morphological analysis.

Performance Results

- Exact match accuracy: 70.00%
- Average edit distance accuracy: 86.69%
- Precision: 0.7000
- Recall: 0.7000
- F-measure: 0.7000
- Average correct reduction ratio: 35.02%
- Average user reduction ratio: 23.67%
- Reduction ratio difference: 11.35%

Observations

Strengths:

1. **Conservative Approach:** The stemmer avoids over-aggressive stemming, which is crucial for maintaining semantic meaning in Nepali.

2. **Good Approximation Performance:** 86.69% edit distance accuracy shows the stems are linguistically reasonable even when not exact matches
3. **Morphological Coverage:** Successfully handles multiple suffix categories (case markers, plurals, verbal suffixes, etc.).
4. **Validation System:** The edit distance and minimum length validation prevents over-stemming.

Weaknesses:

1. **Under-stemming:** 11.35% reduction difference suggests the stemmer is too conservative in some cases.
2. **Limited Sandhi Handling:** Complex morphophonological changes at word boundaries may not be fully addressed.